JURNAL MATEMATIKA DAN KOMPUTER
Vol. 5. No. 3, 178 - 187, Desember 2002, ISSN : 1410-8518

ESTIMASI PARAMETER MODEL TAHAP AWAL AR(1) REGRESI RESPON BINER LONGITUDINAL

Rohmatul Fajriyah
FMIPA UII Yogyakarta
dan
Subanar
FMIPA UGM Yogyakarta

Abstrak

Data yang diperoleh dari hasil pengukuran berulang pada subyek tertentu, biasanya akan berkorelasi. Pada regresi respon biner, jika digunakan model autoregressif order -1, AR(1), maka diperlukan pengetahuan tentang *outcome* sebelumnya, y_0, yang tak terobservasi. Model untuk menginferensi data dengan model AR(1), diantaranya adalah model AR(1) kondisional. Pada model ini, nilai y_0 diambil sembarang, yaitu 0 atau 1. Model di atas akan dibahas dan dibandingkan hasil estimasinya melalui studi simulasi

Kata kunci : Data biner berkorelasi, *Maximum Likelihood*, Dependensi serial.

1. PENDAHULUAN

Dalam percobaan klinis, sering dianalisa data biner yang diperoleh di waktu-waktu yang berurutan untuk menguji hubungan antara probabilitas sukses dan kovariat-kovariat yang bergantung pada waktu.

Data tersebut diperoleh jika diobservasi satu grup pasien per tahun, misalnya, dan observasi-observasi diambil setiap minggu. Tepatnya, setiap subyek atau individu atau pasien mengalami pengukuran berulang mingguan. Jika data diperoleh dari hasil pengukuran berulang per waktu tertentu, maka data akan berkorelasi tinggi . Pada data seperti ini digunakan model autoregressif (AR), khususnya model AR(1). Misal Y_{it} representasi outcome pasien ke $-i$ dalam minggu ke $-t$. Pada $t=1$, maka outcome sebelumnya y_0 tentu tak terobservasi.

Jika diambil $Y_{i0} = 0$, akan timbul masalah dalam pemodelan jika ternyata yang benar adalah $Y_{i0} = 1$, begitu sebaliknya. Permasalahan seperti ini dikenal sebagai permasalahan tahap awal. Penting untuk mengetahui bagaimana mengatasi masalah tahap awal pada regresi data respon biner longitudinal.

Pada tulisan ini, jelasnya, akan dibahas estimasi parameter regresi respon biner longitudinal model AR(1) kondisional, dalam permasalahan tahap awal.

2. DATA LONGITUDINAL

Beberapa penelitian, mengobservasi variabel respon setiap subyek, beberapa kali untuk beberapa waktu tertentu atau pada keadaaan tertentu. Hasil penelitian semacam ini akan menghasilkan data respon berulang.

Jika subyek diobservasi berulang beberapa waktu tertentu, maka data hasil observasi berulang semacam ini disebut sebagai data longitudinal dan studinya disebut studi longitudinal.

Data longitudinal biasanya akan berkorelasi serial dalam subyek. Jelasnya, jika y_{it} merepresentasikan observasi subyek ke-i waktu ke-t, maka subyek i memuat respon berulang y_{it}, yaitu karena observasinya diambil dari subyek yang sama akibatnya respon berulang ini berkorelasi.

Selanjutnya, respon biner dari subyek yang diobservasi beberapa kali beberapa waktu tertentu disebut data respon biner longitudinal.

3. GLM dan MLE

Model-model statistik klasik, untuk menganalisa data regresi, runtun waktu dan longitudinal secara umum berguna dalam situasi-situasi dimana datanya Gaussian dan dapat dijelaskan dengan suatu struktur linear.

Nelder dan Wedderburn pada tahun 1972 memperkenalkan suatu keluarga dari model-model untuk analisis regresi nonstandar dengan respon non normal yang disebut *Generalized Linear Models* (**GLM**)

Maximum likelihood merupakan metode pengestimasian yang sangat populer. Misal $X = \left(X_1, \ldots, X_n\right)$ suatu vektor random observasi-observasi yang distribusi bersamanya adalah suatu fungsi densitas $f_n\left(x|\Theta\right)$ pada ruang Euclide berdimensi - n, R^n. Vektor parameter Θ yang tak diketahui termuat dalam ruang parameter $\Omega \subset R^s$. Untuk X tertentu didefinisikan fungsi likelihood dari X sebagai $L(\Theta) = L_x(\Theta) = f_n\left(x|\Theta\right)$ yang dipandang sebagai fungsi dari $\Theta \subset \Omega$.

4. MODEL AR(1) DALAM PERMASALAHAN TAHAP AWAL

Sebagaimana telah dikemukakan dimuka, data longitudinal merupakan data yang diperoleh dari hasil pengukuran berulang.

Data longitudinal ini dapat dihimpun secara prospektif, mengikuti subyek berkembang sesuai waktu, atau retrospektif, dengan mengekstraksi pengukuran-pengukuran pada setiap subyek dari catatan terdahulunya.

Himpunan data longitudinal pada satu subyek cenderung berinterkorelasi (subyek-subyek biasanya diasumsikan independen),oleh sebab itu diperlukan metode statistik khusus agar diperoleh inferensi yang valid.

4.1. Model AR(1) Kondisional

Diasumsikan data observasi berulang $(y_{it}, x_{it}), t = 1,2,..., n_i$,ada,untuk setiap subyek $i = 1,2,..., m$, dan distribusi bersyarat dari setiap respon y_{it},merupakan fungsi eksplisit dari respon-respon sebelumnya $y_{it-1},..., y_{i1}$ dan kovariat x_{it}, juga probabilitas bersyarat $\Pr\{Y_{it} = 1|Y_{i1},..., Y_{i,t-1}\} = \Pr\{Y_{it} = 1|Y_{i,t-1}\}$ merupakan logit linear . Misal $y_{it-1},..., y_{i1}$ disebut sebagai “history” subyek ke - i pada waktu - t dan dinotasikan dengan H_{it} ,maka $H_{it} = \{y_{ik}, k = 1,..., t-1\}$.

Model yang akan dibahas adalah model dimana distribusi bersyarat dari y_{it} diketahui H_{it} hanya bergantung pada satu observasi sebelumnya, y_{it-1}. Jadi,

$$h\left(\mu_{it}^{'}\right)=x_{it}^{'}\beta+f_1\left(H_{it};\alpha\right) \qquad (4.1.1)$$

atau model ini menyajikan mean bersyarat μ_{it}^{c} sebagai fungsi dari kovariat x_{it} dan respon sebelumnya y_{it-1}. Outcome sebelumnya merupakan variabel penjelas tambahan. Dengan demikian, diperoleh

$$\log it\,\Pr\left(Y_{it}=1\middle|H_{it}\right)=x_{it}^{'}\beta+\alpha y_{it-1}, \qquad (4.1.2)$$

atau

$$\text{logit}\ \Pr\left(Y_{it}=1\middle|H_{it}\right)=x_{it}^{c'}\beta, \qquad (4.1.3)$$

dimana $x_{it}^{c'}$ dan β adalah vektor berukuran $(p+1)$.

Fungsi densitas probabilitas $Y_{it}\middle|Y_{it-1}$ dituliskan sebagai

$$f\left(Y_{it}=y_{it}\middle|Y_{it-1}=y_{it-1}\right)\ =\ \exp\left(y_{it}\eta_{it}^{\ c}-\log\left(1+e^{\eta_{it}^{\ c}}\right)\right) \qquad (4.1.4)$$

Mean dan variansi bersyaratnya ialah

$$\mu_{it}^{c}\ =\ E\left(Y_{it}\middle|H_{it}\right)=a^{'}\left(\theta_{it}\right)=\frac{e^{\eta_{it}^{\ c}}}{1+e^{\eta_{it}^{\ c}}}$$

dan

$$v_{it}^{c}=Var\left(Y_{it}\middle|H_{it}\right)=a^{''}\left(\theta_{it}\right)=\frac{e^{\eta_{it}^{\ c}}}{\left(1+e^{\eta_{it}^{\ c}}\right)^2}.$$

Fungsi likelihood untuk fungsi densitas diatas dituliskan

$$L\left(\beta,Y\right)=\prod_{i=1}^{m}\prod_{t=1}^{n_i}f\left(Y_{it}=y_{it}\middle|Y_{it-1}=y_{it-1}\right) \qquad (4.1.5)$$

dan

*log-likelihood*nya

$$l\left(\beta,Y\right)=\log L\left(\beta,Y\right)=\sum_{i=1}^{m}\sum_{t=1}^{n_i}\log f\left(Y_{it}=y_{it}\middle|Y_{it-1}=y_{it-1}\right) \qquad (4.1.6)$$

Dari persamaan (4.1.4), diperoleh

$$\log f\left(Y_{it}=y_{it}\middle|Y_{it-1}=y_{it-1}\right)\ \ =\ y_{it}\eta_{it}^{\ c}-\log\left(1+e^{\eta_{it}^{\ c}}\right) \qquad (4.1.7)$$

Jadi

$$l\left(\beta,Y\right)\ =\ \log L\left(\beta,Y\right)=\ \sum_{i=1}^{m}\sum_{t=1}^{n_i}y_{it}\eta_{it}^{\ c}-\log\left(1+e^{\eta_{it}^{\ c}}\right) \qquad (4.1.8)$$

Selanjutnya, persamaan (4.1.7) dapat juga dituliskan sebagai

$$\log f\left(Y_{it}=y_{it}\middle|Y_{it-1}=y_{it-1}\right) = y_{it}\log\left(\mu_{it}^{c}\right)+\left(1-y_{it}\right)\log\left(1-\mu_{it}^{c}\right) \qquad (4.1.9)$$

Dengan demikian diperoleh

$$l\left(\beta,Y\right) = \sum_{i=1}^{m}\sum_{t=1}^{n_i} y_{it}\log\left(\mu_{it}^{c}\right)+\left(1-y_{it}\right)\log\left(1-\mu_{it}^{c}\right) \qquad (4.1.10)$$

Persamaan terakhir ini lebih mudah diadaptasi ke dalam bentuk matriksnya.

Namakan $\log f\left(Y_{it}=y_{it}\middle|Y_{it-1}=y_{it-1}\right)=l_{it}\left(\beta\right)$. Akan diperoleh fungsi score sebagai berikut

$$S\left(\beta\right)=\sum_{i=1}^{m}\sum_{t=1}^{n_i}S_{it}\left(\beta\right), \qquad (4.1.11)$$

dimana $S_{it}\left(\beta\right) = \dfrac{\partial l_{it}\left(\beta\right)}{\partial\beta}$.

$S_{it}\left(\beta\right)$ diperoleh dengan menggunakan persamaan (4.1.9), dan mengganti μ_{it}^{c} dengan $h\left(Z_{it}^{c'}\beta\right)$, yaitu

$$S_{it}\left(\beta\right)= \frac{\partial l_{it}\left(\beta\right)}{\partial\beta}$$

$$= \frac{\partial\left(y_{it}\log h\left(Z_{it}^{c'}\beta\right)+\left(1-y_{it}\right)\log\left(1-h\left(Z_{it}^{c'}\beta\right)\right)\right)}{\partial\beta}$$

$$= \frac{y_{it}Z_{it}^{c'}h'\left(Z_{it}^{c'}\beta\right)}{h\left(Z_{it}^{c'}\beta\right)}-\frac{\left(1-y_{it}\right)Z_{it}^{c'}h'\left(Z_{it}^{c'}\beta\right)}{1-h\left(Z_{it}^{c'}\beta\right)}$$

$$= Z_{it}^{c'}h'\left(Z_{it}^{c'}\beta\right)\frac{1}{h\left(Z_{it}^{c'}\beta\right)\left(1-h\left(Z_{it}^{c'}\beta\right)\right)}\left(y_{it}-h\left(Z_{it}^{c'}\beta\right)\right)$$

$$= Z_{it}^{c'}D_{it}\left(\beta\right)\sum\nolimits_{it}^{-1}\left(y_{it}-\mu_{it}^{c}\left(\beta\right)\right) \qquad (4.1.12)$$

dimana

$$D_{it}\left(\beta\right) = h'\left(Z_{it}^{c'}\beta\right)=\frac{\partial h\left(Z_{it}^{c'}\beta\right)}{\partial\beta}=\frac{\partial\mu_{it}^{c}}{\partial\beta}$$

$$\sum_{it}^{-1} = \left(h\left(Z_{it}^{c'}\beta\right)\left(1-h\left(Z_{it}^{c'}\beta\right)\right)\right)^{-1} = \frac{1}{h\left(Z_{it}^{c'}\beta\right)\left(1-h\left(Z_{it}^{c'}\beta\right)\right)}$$

Secara similar akan diperoleh matriks ekspektasi informasi Fisher,

$$G(\beta) = \sum_{i=1}^{m}\sum_{t=1}^{n_i} G_{it}(\beta) \qquad (4.1.13)$$

dimana

$$G_{it}(\beta) = -\frac{\partial l_{it}(\beta)}{\partial\beta\partial\beta'}$$

$$= -\left(-Z_{it}^{c'}h'\left(Z_{it}^{c'}\beta\right)\frac{1}{h\left(Z_{it}^{c'}\beta\right)\left(1-h\left(Z_{it}^{c'}\beta\right)\right)}\left(h'\left(Z_{it}^{c'}\beta\right)\right)'Z_{it}^{c}\right)$$

$$= Z_{it}^{c'}D_{it}(\beta)\sum_{it}^{-1}\left(D_{it}(\beta)\right)'Z_{it}^{c}$$

Selanjutnya MLE untuk β diperoleh dengan metode iterasi pada

$$S(\beta) = \sum_{i=1}^{m}\sum_{t=1}^{n_i} S_{it}(\beta) = 0,$$

dengan mempertimbangkan dua keadaan, yaitu untuk $Y_{i0}=0$ dan $Y_{i0}=1$.

4.2. Contoh Aplikasi

Sebagai aplikasi, disini diambil *data hasil simulasi* untuk model Chan (2000), yang telah dikerjakan Fajriyah (2001).

Data diatas, merupakan data simulasi pasien peserta program MMT (*Methadone Maintenace Treatment*) di Sydney Barat pada tahun 1986. Chan (2000), dalam papernya menyebutkan bahwa, berdasarkan riset doktoralnya, yang dipublikasikan sebagian pada tahun 1998, model bagi pasien MMT, ternyata mengikuti model AR(1). Hal ini mengakibatkan, diperlukannya pengetahuan tentang Y_{i0} untuk setiap pasien, yang tentu saja tak terobservasi.

Model tersebut yaitu:

$$\text{logit}\left\{\Pr\left(Y_{it}=1\middle|Y_{i,t-1}\right)\right\} = \eta_{it} = -0.8423 - 0.00884d_{it} - 0.4049\ln(t) + 2.396Y_{i,t-1} \qquad (4.2.1)$$

dimana, $-0.8423, -0.00884, -0.4049$ dan 2.396 berturut-turut adalah intersep, koefisien slope dosis methadone (dalam miligram), koefisien slope durasi waktu (dalam minggu) dan koefisien slope outcome sebelumnya.

Alasan digunakannya simulasi oleh Fajriyah (2000), adalah tidak dapat diperolehnya data asli MMT.

Berdasarkan perhitungan, untuk m = 10, n = 5, dan ulangan simulasi sebanyak 10 dari *data hasil simulasi*, estimasi parameter simulasi untuk $Y_{i0} = 0$ dan $Y_{i0} = 1$, masing-masing adalah :

Tabel 1. Parameter Hasil Simulasi

Model 4.2.1(0)

Simulasi ke-	β_0	β_1	β_2	β_3
1	-12.18	-0.2558	-1.547	2.407
2	-16.47	-0.2201	-0.4037	2.404
3	-6.085	-0.1281	-0.8805	3.542
4	-11.6	-0.2281	-1.541	2.858
5	-9.342	-0.182	-1.326	3.655
6	-9.892	-0.2235	-0.4015	2.407
7	-2.749	-0.09504	-0.7335	2.411
8	-12.11	-0.2281	-0.9938	2.62
9	-8.812	-0.1559	-0.5466	2.396
10	-9.417	-0.173	-0.4038	2.39

Tabel 2. Parameter Hasil Simulasi

Model 4.2.1(1)

Simulasi Ke-	β_0	β_1	β_2	β_3
1	-11.05	-0.2199	1.004	2.312
2	-11.02	-0.1735	0.7697	3.205
3	-15.43	-0.21	2.43	5.532
4	-9.96	-0.1278	2.644	4.067
5	-12.17	-0.2288	0.595	2.426
6	-7.187	-0.1189	0.387	3.096
7	-11.08	-0.1828	0.7702	2.872
8	-15.41	-0.2565	0.9457	2.881
9	-8.845	-0.2189	2.282	2.31
10	-11.3	-0.1825	0.1183	3.646

Dari kedua tabel hasil simulasi di atas, diperoleh kesimpulan bahwa, untuk sampel kecil (m = 10), pengambilan nilai $Y_{i0} = 0$, ternyata lebih "mendekati" model asli, terutama dari segi interpretasi, dimana hasil ini sejalan dengan kesimpulan Chan (2000), dengan menggunakan sampel asli m = 136, maupun hasil simulasinya sendiri dengan pengulangan simulasi sebanyak 100, m = 136 dan n = 26.

Adapun untuk $Y_{i0} = 1$, meskipun tidak mendekati model asli, namun sejalan dengan hasil simulasi Chan (2000) dengan pengulangan simulasi sebanyak 100, m = 136 dan n = 26.

5. KESIMPULAN

Data respon biner longitudinal, yang diperoleh dari hasil pengukuran berulang pada subyek tertentu, biasanya akan berkorelasi. Jika digunakan model autoregressif order -1, AR(1), maka pengetahuan tentang outcome sebelumnya, y_0, yang tak terobservasi diperlukan untuk inferensi.

Model-model AR(1) yang dapat digunakan untuk mengakomodasi y_0, diantaranya adalah, Model AR(1) Kondisional dan estimasi parameter dilakukan dengan menggunakan metode **MLE**.

DAFTAR PUSTAKA

1. Azzalini, A, *Logistic Regression for Autocorrelated Data with Application to Repeated Measures*, Biometrika, 1994, 81 : 767-775.

2. Chan, J. S. K., *Initial Stage Problem in Autoregressive Binary Regression*, Journal Royal Statistal Society, Part 4, 2000, 49 : 495-502.

3. Chan, J. S. K, and Kuk, A. Y. C, Bell, J and McGilchrist, C, *The Analysis of Methadone Clinic Data Using Marginal and Conditional Logistic Models with Mixture or Random Effects*, Aust. New Zeal. J. Statist, 1998, 40 : 1-10.

4. Diggle, P. J, Liang, K.Y and Zeger, S. L, *Analysis of Longitudinal Data*, Clarendon Press, Oxford, 1994.

5. Fahrmeir, L and Tutz, G, *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer-Verlag, New York, 1994.

6. Fajriyah, R, *Estimasi Parameter Beberapa Model Tahap Awal Regresi Respon Biner Longitudinal*, Tesis S2 Matematika FMIPA UGM, Yogyakarta, 2001.

7. Fitzmaurice, G. M and Laird, N. M, *A Likelihood-Based Method for Analysing Longitudinal Binary responses*, Biometrika, 1993, 80 : 141-151.

8. Liang, K. Y and Zeger, S. L, *A Class of Regression Models for Multivariate Binary Time Series*, J. Am. Statist. Ass, 1989, 84 : 447-451.

9. Ware, J. H, Lipsitz, S and Speizer, F. E, *Issues in the Analyssis of Repeated Categorical Outcome*, Statist. Med, 1988, 31 : 95-108.

10. ---- , *Encyclopedia of Statistical Sciences*, John Wiley and Sons, 1982, Vol. I.